“LAB 5- CLUSTERING”

IENG3304 – DATA MANAGEMENT & ANALYTICS

JUNE 10, 2022

WRITTEN BY: ABILASH SURENDRAN
BANNER ID: B00891410

Table Of Content:

LIST OF FIGURES:

Introduction:

Clustering is a technique for locating comparable groupings of data in a dataset. The black box effect is used to bring a bunch of data together. Clustering is a machine learning technique that uses unsupervised learning. It's like categorization, with the exception that the variable's output isn't predetermined. Several methods of clustering are used in this lab, including k means clustering and hierarchical clustering. The K means clustering method is used to group data by its nearest mean or centroid. Furthermore, the data is divided into k clusters. With the dataset, each cluster creates a pattern that may be utilised to better visualise and comprehend the data. The second step is to perform hierarchical clustering. The data is organised into a flow chart here. The endpoint is a collection of clusters, each of which is unique from the others and contains objects that are broadly comparable to one another. On the wine dataset, both k means clustering and hierarchical clustering are used to uncover different patterns and visualise the data. The report includes the findings of different clustering algorithms.

DATASET:

This set of data pertains to Portuguese red "Vinho Verde" wine variants. The information explains how many active ingredients are present in wine and how they affect its quality. The datasets are used to cluster the data. Alcohol, chlorides, citric acid, density, fixed acidity, free sulphur dioxide, Id, pH, quality, residual sugar, sulphates, total sulphur dioxide, and volatile acidity are some of the columns included. The dataset focuses on how these chemical qualities affect the quality of the wine. The quality of wine tends to alter as minute changes in its chemical composition occur. The amount of alcohol in a wine impacts its flavour and texture, as well as the amount of alcohol that evaporates to carry the wine's scent to our senses. Alcohol also helps to balance sweetness and acidity by adding viscosity. The wine has a saline flavour due to the presence of chloride. Let's have a look at a few chemical components and group them for easier understanding & visualisation.

K means clustering:

K mean clustering aims to find the k data point groups with the least total squared distance between all points and the group's centroid.

One-dimensional clustering:

A wine with high acidity will usually taste sharper and tarter on the palate. A low-acid wine will be smoother and rounder on the palate. Wines with higher acidity levels are more likely to improve over time than those with lower acidity levels because acidity provides some of the necessary backbones for long-term ageing.

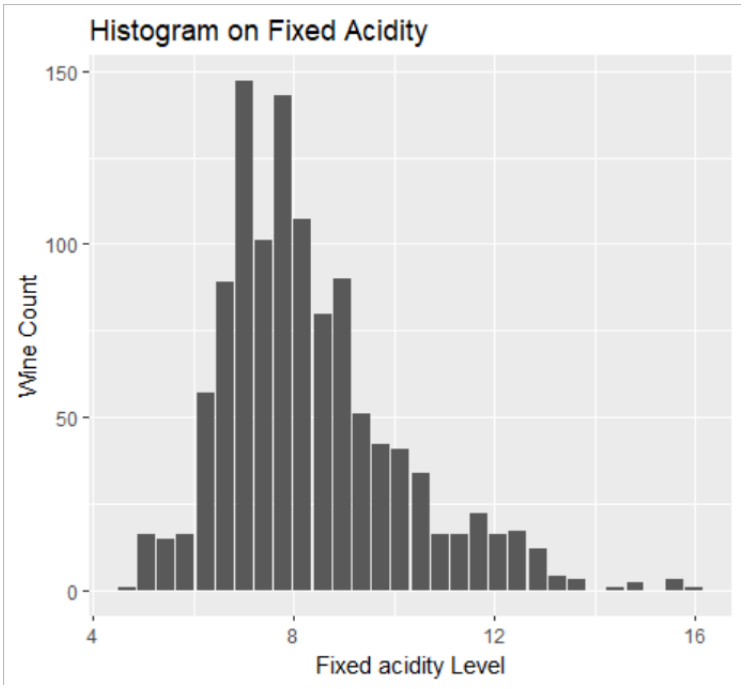

Fig 1: plot for the range of fixed acidity levels in wines

To begin, the fixed acidity level of several wines is presented as a histogram to visualise the trend in which it is spread before clustering. The acidity levels here range from 4 to 16. The taste and quality of wine sharpen as the acidity level rises. For a better understanding, the clustered acidity levels are now presented below.

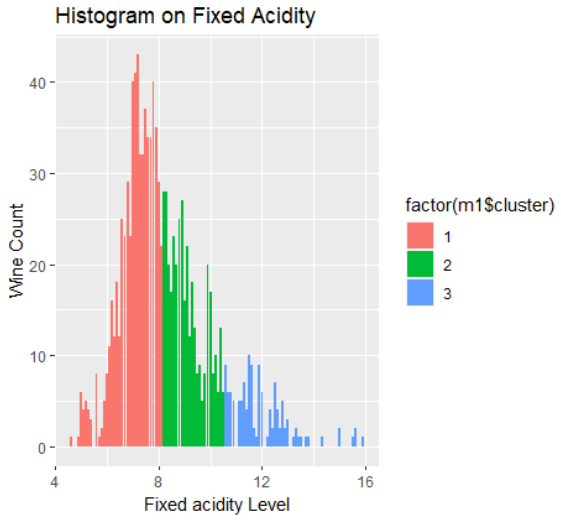

Fig 2: plot with k = 3, for the range of fixed acidity levels in wines

Clustering is used to visualise how the wines' fixed acidity levels are distributed. Because the value of k is kept constant at 3, the data is divided into three clusters. Cluster 1 has the lowest acidity, whereas Cluster 3 has the highest acidity. As a result, the quality of cluster 1 wines will be smoother, while cluster 3 wines will be sharper and tarter.

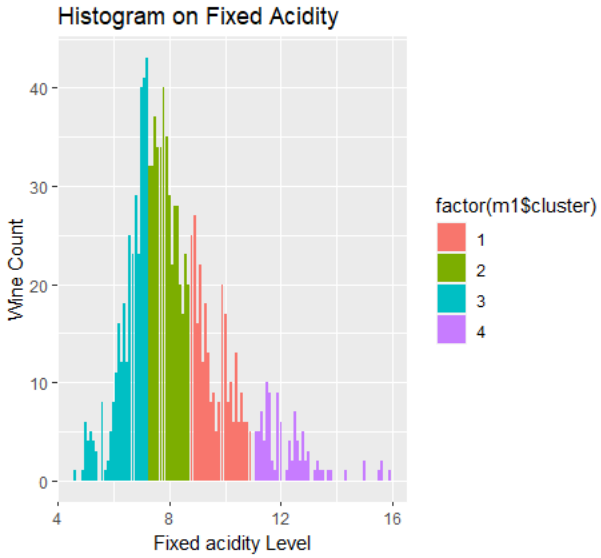

Fig 3: plot with k = 4, for the range of fixed acidity levels in wines

Here there the dataset is clustered into 4 groups. That is cluster 1 has very little acidic content in it. However, the maximum number of wines falls under this category. Here the value of k is 4 and the n-start value is 100.

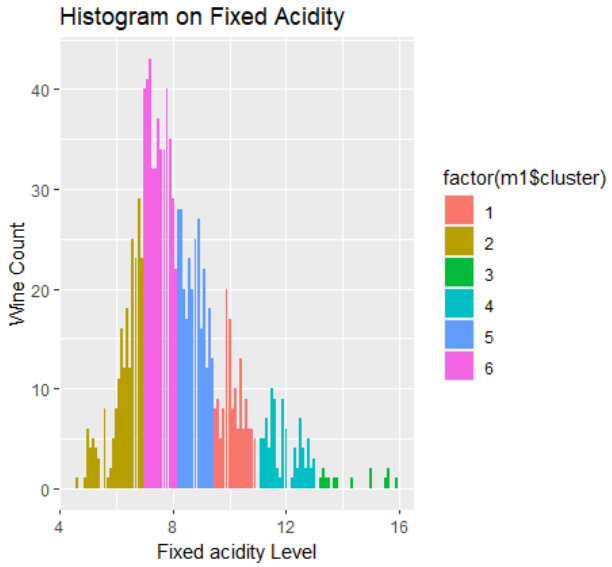

Fig 4: plot with k = 6, for the range of fixed acidity levels in wines

The data is divided into six clusters in this case. The acidity levels are still ranging from 4 to 16. The k value is set to 6 and the n-start value is set to 100. The wine quality in this data set ranges from 3 to 8. As a result, wines in cluster 2 will be classified as low-quality wines, while wines in cluster 3 will be classified as high-quality wines. However, there are additional aspects that can be employed to improve the quality of wines than set acidity levels.

The cluster values with k = 3 make more sense than the rest, because the wine data may be divided into wines of low, moderate, and high grade based on fixed acidity. Furthermore, when the k values rise, the classification of wines by fixed acidity level becomes more complicated.

Two-dimensional clustering:

Relationship between pH and Alcohol

pH stands for "potentially acidic or basic" and refers to the acidic or basic character of a liquid. It is vital to maintain the right pH levels to maintain the wine's quality. The pH of most wines is about 3 or 4; for white wines, a pH of 3.0 to 3.4 is ideal, while for reds, a pH of 3.3 to 3.6 is ideal.

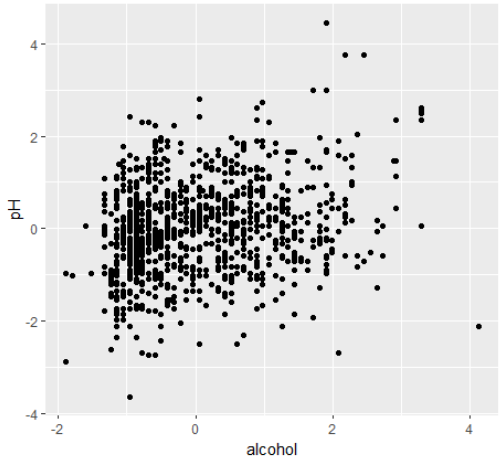

Fig 5: scatter plot for pH and alcohol

A scatter plot for pH and alcohol is made here. The dataset is scaled for improved grouping because the data point value fluctuates in units. Scaling is carried out to improve clustering results. Clustering is done after scaling to find patterns. Furthermore, the mean and variance are retained at 0 and 1, respectively, while scaling.

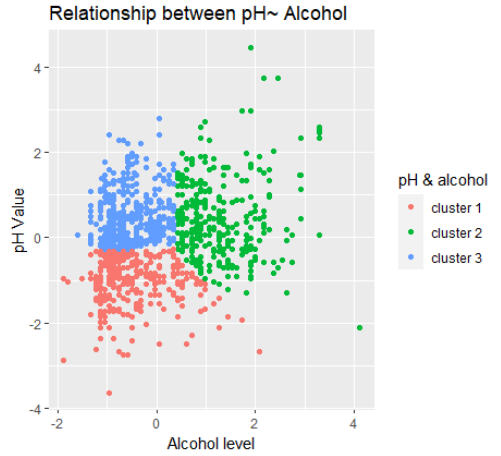

Fig 6: scatter plot for pH and alcohol with k=3

The data is divided into three clusters, as can be seen in the graph. That is, the value of k is preserved at 3 in the k means technique, and the equivalent result is obtained. Clusters 1, 2 and 3 as 450, 375 and 318 points respectively.

```
> M2$centers
     alcohol         pH
1 -0.4992939 -1.0064947
2  1.2985827  0.4162508
3 -0.5015869  0.5445950
```

The centre points of these three clusters are listed above. The K-means algorithm helps in identifying the centre points.

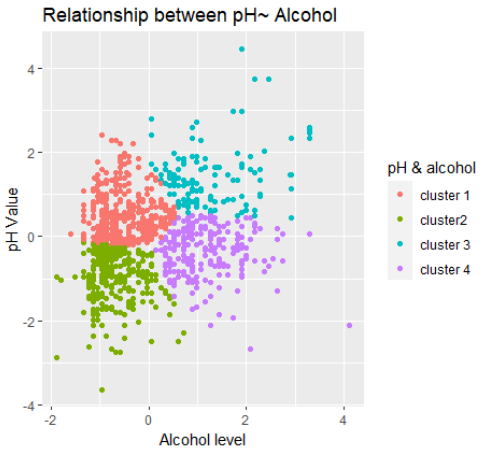

Fig 7: scatter plot for pH and alcohol with k=4

The data is divided into four clusters, as can be seen in the graph. That is, the value of k is preserved at 4 in the k means technique, and the equivalent result is obtained. According to the report, clusters 2 and 4 have lower pH values than clusters 1 and 3.

	alcohol	pH
1	1.1364703	-0.3745518
2	-0.4895011	0.5233914
3	1.2779712	1.4384203
4	-0.7019819	-0.9823979

The centre points of these four clusters are listed above. The K-means algorithm helps in identifying the centre points.

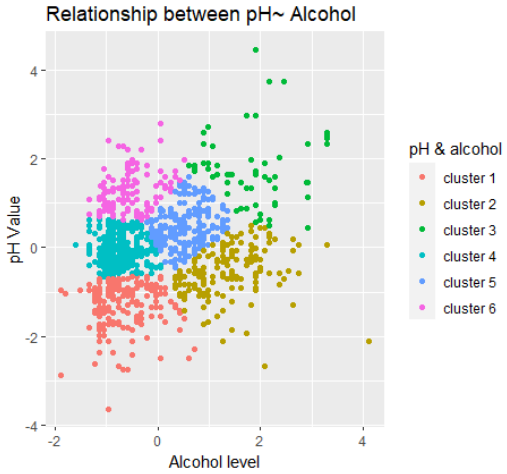

Fig 8: scatter plot for pH and alcohol with k=6

The value of k is set to 6 in this case, dividing the plot into six distinct clusters. Cluster 1 has 137 data points, Cluster 2 has 63 data points, Cluster 3 has 238 data points, Cluster 4 has 306 data points, Cluster 5 has 224 data points, and Cluster 6 has 175 data points. However, multiple clusters overlap the same pH level, making clear interpretation difficult.

```
> M2$centers
     alcohol          pH
1  1.8827633  1.77673301
2  1.3065000 -0.56490677
3  0.4701543  0.45669793
4 -0.6735080 -1.25972836
5 -0.7485571 -0.03755906
6 -0.5782778  1.25476365
```

The centre points of these six clusters are listed above. The K-means algorithm helps in identifying the centre points.

DENSITY~RESIDUAL SUGAR:

Wines have a density that is often higher than that of water. This is because the presence of residual sugar in the wine increases the density of the wine and thus enhances the quality of the wine.

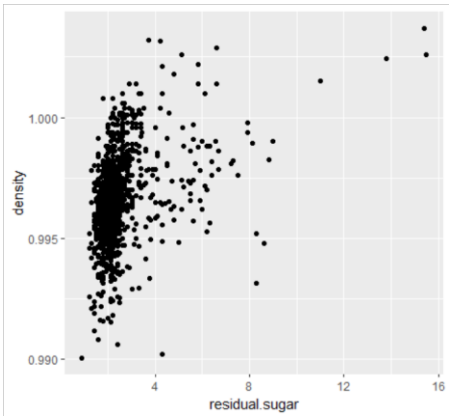

Fig 9: scatter plot for residual sugar and density

To better comprehend the link between density and residual sugar, a scatter plot is constructed. The graph has a range of 0 to 16. The majority of the points are plotted within the range of 4, while the remainder is dispersed irregularly across the scale.

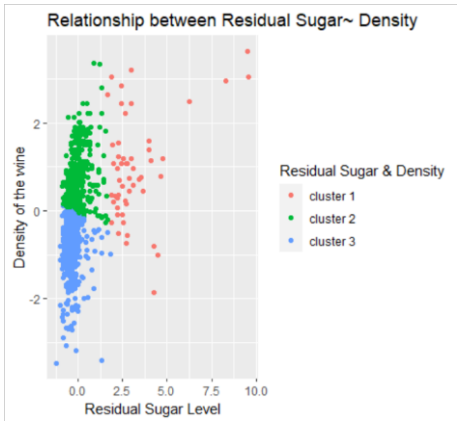

Fig 10: scatter plot for residual sugar and density with k = 3

The dataset is initially clustered into three clusters using a K value of three. To improve clustering accuracy, the data is scaled. The left-hand huge clump is broken into two clusters, while the right-hand outliers are separated into a third cluster. The density of the wine increases as the residual sugar level rises. The centre points of these three clusters are listed below for

the scatter plot for density and residual sugar. The K-means algorithm helps in identifying the centre points.

```
> M3$centers
     density residual.sugar
1  0.9632268     3.28182144
2 -0.7484315    -0.34111364
3  0.7112771    -0.03186697
```

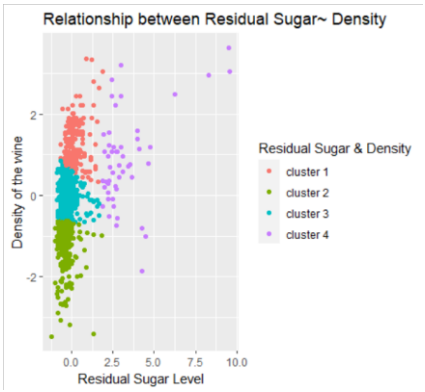

Fig 11: scatter plot for residual sugar and density with k = 4

The dataset is initially clustered into four clusters using a K value of four. To improve clustering accuracy, the data is scaled. The majority of the points in the dataset are on the right side of the plot. This large section is broken into three distinct groups. That is, even when the residual sugar level of the wines is the same, the density of the wines varies. This explains why residual sugar isn't the sole component that influences the density of a wine. The centre points of these four clusters are listed below for the scatter plot for density and residual sugar. The K-means algorithm helps in identifying the centre points.

```
> M3$centers
       density residual.sugar
1 -0.001821677     -0.2577912
2  0.902645570      3.3301646
3 -1.222538537     -0.3660701
4  1.249776983      0.1909088
```

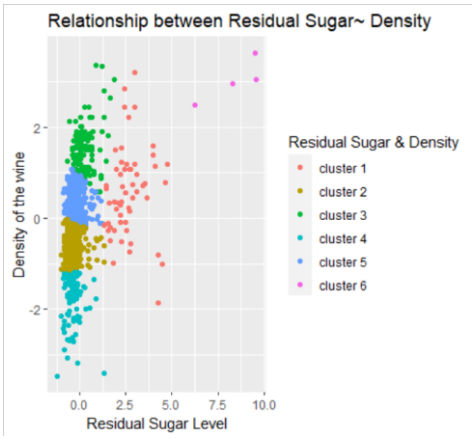

Fig 12: scatter plot for residual sugar and density with k = 6

Clustering is produced as a result of another variation in k. For a better study of the dataset, the k value has been increased to 6. Cluster 6 is the outlier here. Cluster 6 wines have a higher density than the others because their residual sugar level is more. The centre points of these six

clusters are listed below for the scatter plot for density and residual sugar. The K-means algorithm helps in identifying the centre points.

	density	residual.sugar
1	2.9217624	8.0274166
2	1.5155492	0.2553787
3	0.5595653	2.6164561
4	-1.7845197	-0.3750345
5	-0.5442256	-0.3397241
6	0.3678477	-0.1964154

pH~ Chlorides

Chloride is a base, and it thereby decreases the quality of the wine. Sulphur dioxide molecules and sulphite ions make up the majority of sulphates present in wine. Many experts say that a higher sulphurous content in wine results in a duller flavour and that the high intensity of sulphite ions poses a health danger while also speeding up the fermentation process.

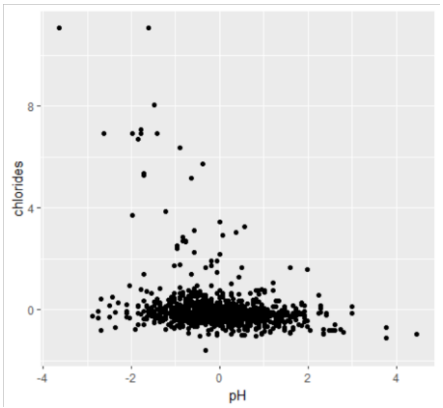

Fig 13: scatter plot for pH and Chlorides

The data is scaled and the scatter plot for establishing the relationship between the pH and Chlorides is created.

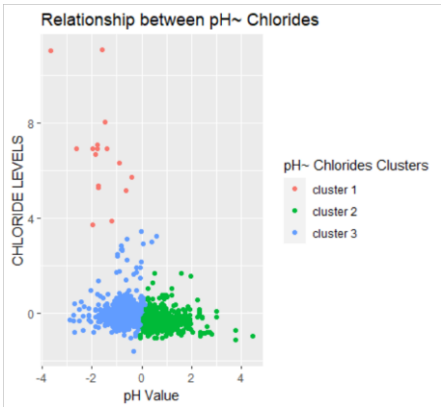

Fig 14: scatter plot between pH~ Chlorides with k = 3

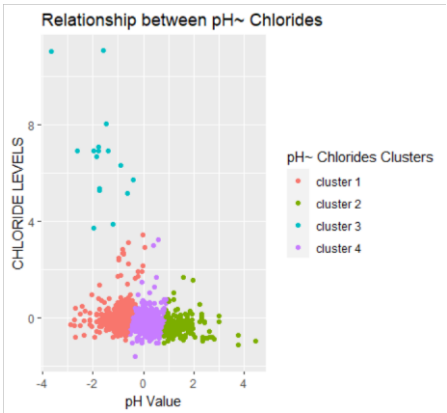

Fig 15: scatter plot between pH~ Chlorides with k = 4

The k value is set to 3 and 4, and the resulting graph is produced. The graph clearly shows how the chloride levels affect the wine's total pH value. When the chloride level in the wine

is high, the pH drops, whereas when the chloride level drops, the pH rises. As it creates a change in the pH value, it thereby affects the overall quality of the wine.

```
> M3$centers
          pH   chlorides
1 -1.692363  6.71542595
2  1.482805 -0.27959180
3  0.175945 -0.15895900
4 -1.022001  0.08639803
```

The centre points of the four clusters in fig 15 are listed above. The K-means algorithm helps in identifying the centre points.

```
> M3$centers
          pH   chlorides
1 -0.7069751  0.01858974
2 -1.6923634  6.71542595
3  0.8120380 -0.23020268
```

The centre points of the three clusters in fig 14 are listed above. The K-means algorithm helps in identifying the centre points.

Hierarchical clustering:

Hierarchical clustering, also known as hierarchical cluster analysis, is a method of grouping related objects into clusters. The endpoint is a collection of clusters, each of which is distinct from the others yet the items within each cluster are broadly similar[3]. There are several methods like complete, average and single. Here the complete method is used.

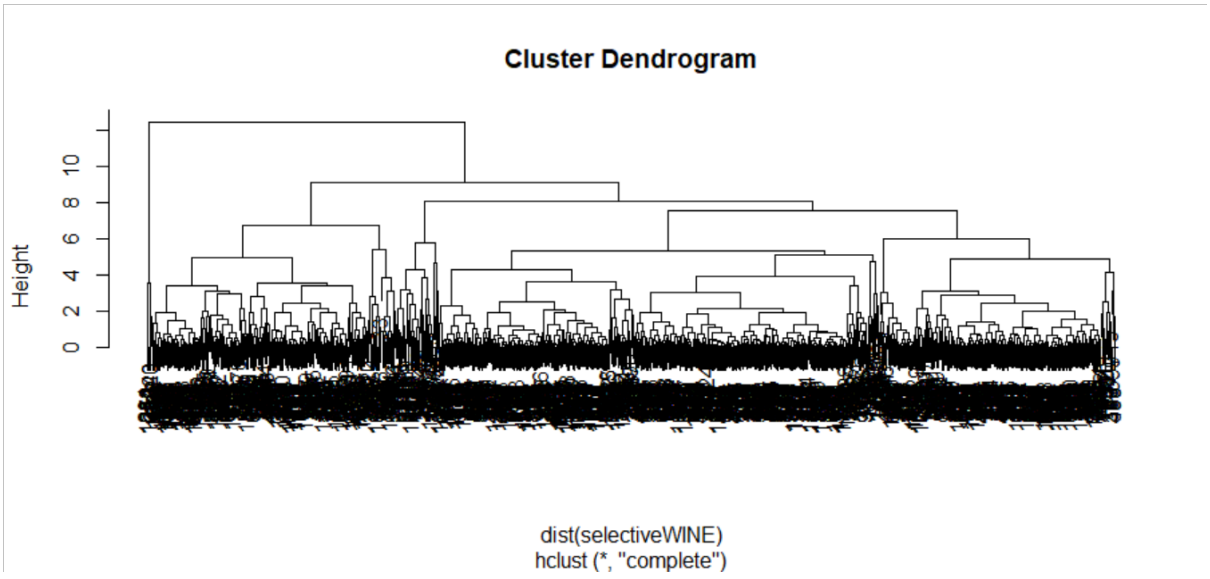

Fig 16: CLUSTER DENDROGRAM

When raw data is provided, the system calculates the distance matrix automatically. However, if the dataset contains many data points, it can be difficult to interpret the dendrogram's result. The dendrogram, on the other hand, can be sliced at any height and the number of clusters obtained can be viewed at that specific height. For example, if we cut the above dendrogram at h=9, we obtain two clusters, however, if we cut at h=7, we get four clusters.

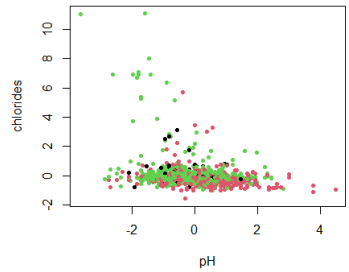

Fig 17: Hierarchical clustering for pH ~chlorides

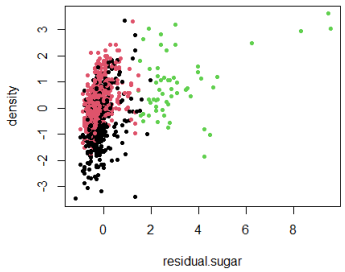

Fig 18: Hierarchical clustering for residual sugar~ density

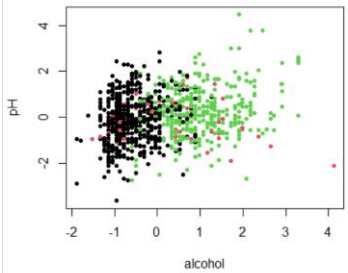

Fig 19: Hierarchical clustering for alcohol ~pH

The data is clustered based on various dimensions using the multi-dimensional clustering technique. With a k value of 3, hierarchical clustering is performed. pH ~ chlorides, residual sugar ~ density, and alcohol ~ pH are all tested. The plots that were collected are listed below. The plots are difficult to comprehend, as the data points are crowded one over the other, as shown in the graph.

Conclusion:

In this lab report, the clustering technique is understood in detail. A dataset about red wines with 1143 data points was taken and clustered. The dataset was well explored with several k and nstart values and the corresponding graphs were recorded. Both k means and hierarchical clustering were performed, and the corresponding graphs were recorded. Moreover, k means clustering gives better results, as the data can be interpreted easily.

R- CODE:

```
library(tidyverse)

library(dplyr)

library(ggplot2)

library(stats)

install.packages("factoextra")

#importing dataset

setwd("C:/CANADA/industrial engineering/summer term/data analytics/dataset")

wine= read.csv("WineQT.csv")

head(wine)

str(wine)

dim(wine)

view(wine)
```

```
#cleaning dataset
colSums(is.na(wine))
cleanwine <- wine[complete.cases(wine),]
dim(cleanwine)
cleanwine$quality <- as.factor(cleanwine$quality)

library(GGally)
ggpairs(cleanwine, aes(colour = quality))

#1D datsaset
attach(cleanwine)
ggplot(data= cleanwine)+
  geom_histogram(aes(x=fixed.acidity),
                 position = position_dodge2(padding = .1,
                                            preserve = "single"))+
  ylab("Wine Count")+
  xlab("Fixed acidity Level")+
  ggtitle("Histogram on Fixed Acidity ")

#creating 1 d dataset
Selectivewine = dplyr::select(cleanwine, fixed.acidity)
m1 = kmeans(Selectivewine, 4, nstart=100)
m1$size
#histogram for fixed acidity
#k value is 4
ggplot(data = Selectivewine, aes(x = fixed.acidity, fill = factor(m1$cluster)))+
  geom_histogram(stat = "count")+xlab("Fixed acidity Level")+
  ylab("Wine Count")+ labs(colour = "Cluster Number")+
  ggtitle("Histogram on Fixed Acidity")+
  scale_colour_discrete(name = "Fixed Acidity Clusters",
```

```
                labels = c("cluster3", "cluster 2"
                        ,"cluster 1", "cluster 4"))
#k value is 3
m1 = kmeans(Selectivewine, 3, nstart=10)
m1$size
#histogram for fixed acidity
ggplot(data = Selectivewine, aes(x = fixed.acidity, fill = factor(m1$cluster)))+
  geom_histogram(stat = "count")+xlab("Fixed acidity Level")+
  ylab("Wine Count")+ labs(colour = "Cluster Number")+
  ggtitle("Histogram on Fixed Acidity")+scale_colour_discrete(name = "Fixed Acidity Clusters",
                labels = c("less acidic", "moderately acidic"
                        ,"highly acidic"))
#k value is 6
m1 = kmeans(Selectivewine, 6, nstart=10)
m1$size
#histogram for fixed acidity
ggplot(data = Selectivewine, aes(x = fixed.acidity, fill = factor(m1$cluster)))+
  geom_histogram(stat = "count")+xlab("Fixed acidity Level")+
  ylab("Wine Count")+ labs(colour = "Cluster Number")+
  ggtitle("Histogram on Fixed Acidity")+scale_colour_discrete(name = "Fixed Acidity Clusters",
                                        labels = c("less acidic", "moderately acidic"
                                                ,"highly acidic"))
#2 dimensional
winescale = scale(wine)
winescale = data.frame(winescale)
attach(winescale)
ggplot(mapping = aes(x = alcohol, y = pH))+
  geom_point()
#k 6 pH~alcohol
```

```
attach(winescale)
Selectivewine1 = dplyr::select(winescale, alcohol,pH)
M2 = kmeans(Selectivewine1, 6, nstart = 3)
M2$size
M2$centers
ggplot(data = Selectivewine1, aes(x=alcohol, y = pH, colour = factor(M2$cluster)))+
  geom_point()+
  xlab("Alcohol level")+
  ylab("pH Value")+
  labs(colour = "heart rate range clusters")+
  ggtitle("Relationship between pH~ Alcohol")+
  scale_colour_discrete(name = "pH & alcohol",
                        labels = c("cluster 1", "cluster 2"
                                   ,"cluster 3", "cluster 4", "cluster 5" , "cluster 6"))
#k3 pH~alcohol
Selectivewine1 = dplyr::select(winescale, alcohol,pH)
M2 = kmeans(Selectivewine1, 3, nstart = 10)
M2$size
M2$centers
?kmeans
ggplot(data = Selectivewine1, aes(x=alcohol, y = pH, colour = factor(M2$cluster)))+
  geom_point()+
  xlab("Alcohol level")+
  ylab("pH Value")+
  labs(colour = "heart rate range clusters")+
  ggtitle("Relationship between pH~ Alcohol")+
  scale_colour_discrete(name = "pH & alcohol",
                        labels = c("cluster 1", "cluster 2"
                                   ,"cluster 3"))
#k4 pH~alcohol
```

```
Selectivewine1 = dplyr::select(winescale, alcohol,pH)
M2 = kmeans(Selectivewine1, 4, nstart = 100)
M2$centers
ggplot(data = Selectivewine1, aes(x=alcohol, y = pH, colour = factor(M2$cluster)))+
  geom_point()+
  xlab("Alcohol level")+
  ylab("pH Value")+
  labs(colour = "heart rate range clusters")+
  ggtitle("Relationship between pH~ Alcohol")+
  scale_colour_discrete(name = "pH & alcohol",
                        labels = c("cluster 1", "cluster2"
                                   ,"cluster 3","cluster 4"))

#clustering using 2 dimensional
#density~residual sugar
attach(winescale)
ggplot(mapping = aes(x = residual.sugar, y = density))+
  geom_point()
#k4
attach(winescale)
Selectivewine2 = dplyr::select(winescale, density, residual.sugar)
M3 = kmeans(Selectivewine2, 4, nstart = 100)
M3$centers
ggplot(data = Selectivewine2, aes(x = residual.sugar, y = density,
                                  colour = factor(M3$cluster)))+
  geom_point()+
  xlab("Residual Sugar Level")+
  ylab("Density of the wine")+
  labs(colour = "heart rate range clusters")+
  ggtitle("Relationship between Residual Sugar~ Density")+
```

```
  scale_colour_discrete(name = "Residual Sugar & Density ",
                        labels = c("cluster 1","cluster 2","cluster 3"
                                   ,"cluster 4"))
#k3
attach(winescale)
Selectivewine2 = dplyr::select(winescale, density, residual.sugar)
M3 = kmeans(Selectivewine2, 3, nstart = 10)
M3$centers
ggplot(data = Selectivewine2, aes(x = residual.sugar, y = density,
                                  colour = factor(M3$cluster)))+
  geom_point()+
  xlab("Residual Sugar Level")+
  ylab("Density of the wine")+
  labs(colour = "heart rate range clusters")+
  ggtitle("Relationship between Residual Sugar~ Density")+
  scale_colour_discrete(name = "Residual Sugar & Density ",
                        labels = c("cluster 1","cluster 2","cluster 3"
                                   ,"cluster 4"))
#k6 does not make sense
attach(winescale)
Selectivewine2 = dplyr::select(winescale, density, residual.sugar)
M3 = kmeans(Selectivewine2, 6, nstart = 10)
M3$size
M3$centers
ggplot(data = Selectivewine2, aes(x = residual.sugar, y = density,
                                  colour = factor(M3$cluster)))+
  geom_point()+
  xlab("Residual Sugar Level")+
  ylab("Density of the wine")+
  labs(colour = "heart rate range clusters")+
```

```
  ggtitle("Relationship between Residual Sugar~ Density")+
  scale_colour_discrete(name = "Residual Sugar & Density ",
                        labels = c("cluster 1","cluster 2","cluster 3",
                                   "cluster 4","cluster 5","cluster 6"))



#clustering using 2 dimensional
#ph~ chlorides

winescale = scale(wine)
winescale = data.frame(winescale)
attach(winescale)
ggplot(mapping = aes(y = chlorides, x = pH))+
  geom_point()

#k4
attach(winescale)
Selectivewine3 = dplyr::select(winescale, pH,chlorides)
M3 = kmeans(Selectivewine3, 4, nstart = 100)
M3$centers
ggplot(data = Selectivewine3, aes(x = pH, y = chlorides,
                                  colour = factor(M3$cluster)))+
  geom_point()+
  xlab("pH Value")+
  ylab("CHLORIDE LEVELS")+
  labs(colour = "heart rate range clusters")+
  ggtitle("Relationship between pH~ Chlorides")+
  scale_colour_discrete(name = "pH~ Chlorides Clusters ",
                        labels = c("cluster 1","cluster 2","cluster 3"
                                   ,"cluster 4"))
```

```
#k3
attach(winescale)
Selectivewine3 = dplyr::select(winescale, pH,chlorides)
M3 = kmeans(Selectivewine3, 3, nstart = 10)
M3$centers
ggplot(data = Selectivewine3, aes(x = pH, y = chlorides,
                                  colour = factor(M3$cluster)))+
  geom_point()+
  xlab("pH Value")+
  ylab("CHLORIDE LEVELS")+
  labs(colour = "heart rate range clusters")+
  ggtitle("Relationship between pH~ Chlorides")+
  scale_colour_discrete(name = "pH~ Chlorides Clusters ",
                        labels = c("cluster 1","cluster 2","cluster 3"
                                   ,"cluster 4"))



#hierarchical clustering
library(factoextra)
#scale data set
scaledWine<- scale(wine)
attach(wine)
selectiveWINE = dplyr::select(winescale, pH, alcohol, residual.sugar, quality)
selectiveWINE= data.frame(selectiveWINE)
view(scaledWine)
m2h = hclust(dist(selectiveWINE), method="complete")
plot(m2h)
attach(selectiveWINE)
par(mar=c(1, 1, 1, 1))
plot(pH~density, col=(cutree(m2h, k=2)))
```

```
plot(pH~density, col=(cutree(m2h, h=5)))
hm1 = hclust(dist(selectiveWINE), method="average")
plot(hm1)

attach(selectiveWINE)
SF = scale(selectiveWINE)
SF = data.frame(SF)
m1 = kmeans(SF, 3, nstart = 10)
plot(chlorides~pH, col=m1$cluster, pch=20)
m1 = kmeans(SF, 3, nstart = 10)
plot(density~residual.sugar, col=m1$cluster, pch=20)
m1$tot.withinss
m1 = kmeans(SF, 3, nstart=100)
plot(pH~alcohol, col=m1$cluster, pch=20)
```

REFERENCE:

1. https://www.kaggle.com/datasets/yasserh/wine-quality-dataset?datasetId=1866301&searchQuery=clu
2. Professor Scott Flemming's code on clustering. – "Chapter 10 in-class exercise unsupervised learning K-means clustering and Hierarchical clustering" - https://dal.brightspace.com/d2l/le/content/221958/viewContent/3009644/View.
3. https://www.coursehero.com/file/p6st6ffh/highest-CCC-is-3-if-the-Average-Method-is-used-instead-of-the-default-Ward/